Image Retrieval Based on Fused CNN Features

Feng ZHANG* and Bao-jiang ZHONG
School of Computer Science and Technology, Soochow University,
Suzhou, Jiangsu 215000, China
*E-mail: 20144227006@stu.suda.edu.cn

Keywords: Convolutional neural networks, Feature extraction, Image retrieval, Feature fusion.

Abstract. Recently, various convolutional neural network (CNN) models have demonstrated their powerful ability as a universal representation for different image recognition tasks. In this paper, image retrieval with different kinds of CNN is considered. Firstly, we extract three kinds of CNN features with the current popular pre-trained CNN models to process image retrieval, respectively. Then, we compute weighted average of the similarity scores of these CNN features and propose an image retrieval algorithm based on the fused CNN features. Extensive experiments on two publicly available datasets well demonstrate that the proposed algorithm is clearly better than the retrieval algorithms based on individual CNN features and other current image retrieval algorithms.

Introduction
The universality of smartphones, digital cameras and other portable image capture devices, as well as the popularity of WeChat, micro-blog and other data sharing platforms, have led to an explosion of images on the network. Faced with the massive image data, the methods of traditional manual annotation have been unable to cope with it. The content-based image retrieval has gradually become a hot spot of scientific research, and the related application of image retrieval technology has been developed.

In recent years, due to the development of convolutional neural networks (CNN) [1, 2], significant progress has been made in visual recognition, e.g., image classification and object detection. Especially, a big breakthrough in image classification is made by [2], which has achieved the state-of-the-art performance (with 10% gain over the method based on hand-crafted features) in large-scale object recognition. The CNN architecture combines low-level features into abstract high-level features with non-linear transformation, which enables it to have capacity to learn the semantic representation from images. So far, many CNN models have been proposed, such as the AlexNet model [2], the VGGNet model [3], the GoogLeNet model [4], the ResNet model [5] and so on. Donahue et al. [6] and Razavian et al. [7] demonstrated that features extracted from the pre-trained CNN models can be considered as a generic image representation for diverse visual recognition tasks. A separate but related to the image classification problem is the issue of image retrieval, i.e. the task of finding images containing the similar object or scene as in a query image. It has been suggested that the feature of the upper layer can be considered as a good descriptor for image retrieval [2, 6]. To the best of our knowledge, some previous articles [7, 8, 9] have already attempted to apply CNN features to image retrieval and obtained good retrieval results. However, they only use information from individual CNN features which cannot comprehensively describe the information of the image. In order to improve the accuracy of the retrieval algorithm, we combine three kinds of existing CNN models, and propose an image retrieval algorithm based on fused CNN features. The main idea of the new algorithm is as follows. Firstly, the VGGNet CNN feature, the GoogLeNet CNN feature and the ResNet CNN feature of the query image are extracted. Then the weighted average similarities between the query features and the features of database images need to be computed. Finally, the images with top-n highest scores are returned as the result.
CNN Models and Feature Extraction

Through an experimental comparison, we choose the pre-trained VGGNet model [3], GoogLeNet model [4] and ResNet model [5] to extract different kinds of CNN features. These CNN models and CNN feature extraction process are mainly introduced as following.

**VGGNet model.** VGG increases the depth of the original architecture [2] to 16-19 layers, which is substantially deeper than what has been used in the prior art. To reduce the number of parameters in such deep networks, VGG uses very small (3×3) convolution filters in all layers and the convolutional stride is set to 1 pixel. Considering the speed and accuracy, we choose the 16-layer pre-trained VGGNet model to extract image features. For convolution layers and max-pooling layers have much higher dimension, fully-connected layers are usually used as feature representation for images. Following Hiton’s guidance like AlexNet [2], we extract 4096 dimensional feature of the layer fc7 as the description of each input image. The structure of VGGNet-16 model can be found in [3].

**GoogLeNet model.** GoogLeNet is a 22-layer deep convolutional neural network architecture. The main hallmark of this architecture is the improved utilization of the computing resources inside the network. To optimize its quality, the architectural decisions are based on the Hebbian principle and the intuition of multi-scale processing. It uses 12 times fewer parameters than AlexNet [2], while being significantly more accurate. Its architectural details can be found in [4]. According to our experience, we extract 1024 dimensional feature of the layer pool5 as the representation of the image for image retrieval.

**ResNet model.** ResNet is a residual learning framework that is deeper but needs less training time than those used previously. These residual networks are easier to optimize, and can gain accuracy from considerably increased depth. In the ImageNet dataset, the depth of these residual nets is up to 152 layers; that is 7 times deeper than VGGNet [3], but still have lower complexity. The architectures of residual networks can been seen in [5]. In our experiments, we choose the 152-layer pre-trained model to extract the 2048 dimensional feature of the layer pool5 as the representation of the image.

It is known that normalization and PCA are common methods used in the feature processing step [7]. Accordingly, when all the CNN features are extracted, we reduce the dimensionality of the feature vectors to 128 by PCA in bid to speed up the retrieval time, and each feature vector is normalized in l2-norm which is convenient to measure the similarity of two feature vectors by cosine distance. Retrieval accuracy of the individual CNN features from the above three pre-trained models on two datasets is presented in Table 1.

![Table 1. Retrieval accuracy of the individual CNN features on two datasets.](image)

<table>
<thead>
<tr>
<th>Datasets/CNN models</th>
<th>VGGNet-16</th>
<th>GoogLeNet</th>
<th>ResNet-152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel-10k, mAP</td>
<td>0.552</td>
<td>0.517</td>
<td>0.617</td>
</tr>
<tr>
<td>Caltech-256, mAP</td>
<td>0.467</td>
<td>0.49</td>
<td>0.469</td>
</tr>
</tbody>
</table>

**Image Similarity Calculation**

In this paper, we use the weighted feature distance to compute the similarity between the query image denoted as $Q$ and each image denoted as $I$ in the database. Then the result of retrieval is returned according to the order of similarity. Here the cosine distance is used to measure the similarity of two feature vectors by cosine distance. Retrieval accuracy of the individual CNN features from the above three pre-trained models on two datasets is presented in Table 1.

$$S(Q, I) = \frac{h(Q) \cdot h(I)}{\|h(Q)\| \|h(I)\|}$$

(1)

where $h(Q)$ and $h(I)$ represent the CNN features of $Q$ and $I$, respectively, and both $\|h(Q)\|$ and $\|h(I)\|$ are equal to 1 due to the CNN features are l2-normalized.
The fusion strategy of multiple feature similarity scores includes the sum, product, maximum, minimum rules and so on. Among them, we employ the product rule for the following two reasons. Firstly, previous work [10, 11, 12] proved that the product rule has the performance of not less than the sum rule. Secondly, when using product rule, the feature combination of different meaning does not need heavy feature normalization. Based on the above three kinds of CNN feature similarity scores and product rule, the similarity between $Q$ and $I$ is defined as below.

$$Sim(Q, I) = S_v(Q, I)^{w_v} \cdot S_g(Q, I)^{w_g} \cdot S_r(Q, I)^{w_r},$$

where $w_v$, $w_g$, $w_r$ are the weights corresponding to the above three CNN feature similarity scores with $w_v + w_g + w_r = 1$. Based on the results of simulation experiments, the default values of these parameters in Corel-10k dataset are taken as $w_v = 0.2$, $w_g = 0.2$, $w_r = 0.6$ and these in Caltech-256 dataset are taken as $w_v = 0.3$, $w_g = 0.4$, $w_r = 0.3$. Note that, (2) can be transformed into a sum form by a logarithmic operation to reduce computational complexity. The flow chart of our proposed image retrieval algorithm is shown in Fig. 1.

![Flow chart of the image retrieval based on fused CNN features.](image)

**Experiments**

**Datasets and Evaluation Criteria**

Two datasets are adopted to test our retrieval algorithm, including Corel-10k [13] and Caltech-256 [14]. Corel-10k dataset contains 10000 images, a total of 100 categories, each category contains 100 images. Images of the same category contain the same semantic information, including the categories of flowers, birds, beaches, buildings, etc.. Each image in the dataset is used as query. Caltech-256 dataset contains 30607 images consisting of 256 categories, including furniture, animals, sports equipment, vehicles and so on. We randomly select 5000 images from the database as queries to test the performance of the algorithm. In all experiments, Precision-Recall curve (PR curve) and Mean Average Precision (mAP) are used to evaluate our algorithm on the above two datasets.

**Comparison of Retrieval Performance of Individual Features and Fused Features**

An individual CNN features may work well on some samples. However, it will be worse in others. To overcome this difficulty, we use the fused CNN features for image retrieval. In order to study the contribution of different CNN features, we compare the retrieval effectiveness of the three individual CNN features as well as their fused form. Because it is difficult to automatically adjust the weights of different features, here we can only manually adjust them. Fig. 2 illustrates the PR curves of retrieval...
performance of the three individual CNN features and their fused form on the Corel-10k dataset and Caltech-256 dataset. In the legends, V, G and R represent VGGNet-16, GoogLeNet and ResNet-152, respectively. An important conclusion from Fig. 2 can be drawn that the retrieval results of the fused features are better than that of the individual features.

![PR curves of retrieval performance of individual features and fused features.](image)

**Figure 2.** PR curves of retrieval performance of individual features and fused features. (a) the Corel-10k dataset; (b) the Caltech-256 dataset.

**Comparison between Our Algorithm and the Existing Retrieval Algorithms**

According to the above experimental results, we choose CNN features of VGGNet-16, GoogLeNet and ResNet-152 to describe the semantics of the image, and complete the image retrieval based on fused features by the weighted similarity calculation. In order to fairly evaluate the retrieval accuracy of our algorithm, we select the following retrieval algorithms as references, MSD [13], CDH [15], HE [16], PCA-CNN [9] and we also extract the VGGNet-16 CNN features for image retrieval which acts as the CNN-based benchmark algorithm. For the sake of fairness, CNN features of the PCA-CNN algorithm are extracted from the pre-trained VGGNet-16 which is better than the original CNN model.

The PR curves of the comparison algorithms and our algorithm on Corel-10k dataset and Caltech-256 dataset are plotted in Fig. 3(a) and 3(b). Table 2 shows the specific numerical results on the two datasets. It can be seen from Fig. 3 and Table 2 that the proposed algorithm is clearly higher than the algorithms of MSD, CDH, HE, PCA-CNN and CNN-based benchmark.

![PR curves of other algorithms and our algorithm.](image)

**Figure 3.** The PR curves of other algorithms and our algorithm. (a) the Corel-10k dataset; (b) the Caltech-256 dataset.
Table 2. Retrieval results of other algorithms and our algorithm.

<table>
<thead>
<tr>
<th>Datasets/Algorithms</th>
<th>MSD</th>
<th>CDH</th>
<th>HE</th>
<th>Benchmark</th>
<th>PCA-CNN</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel-10k, mAP</td>
<td>0.179</td>
<td>0.18</td>
<td>0.1</td>
<td>0.519</td>
<td>0.552</td>
<td><strong>0.644</strong></td>
</tr>
<tr>
<td>Caltech-256, mAP</td>
<td>0.03</td>
<td>0.039</td>
<td>0.02</td>
<td>0.45</td>
<td>0.467</td>
<td><strong>0.536</strong></td>
</tr>
</tbody>
</table>

Figure 4. An example of image retrieval using our algorithm on the Caltech-256 dataset. (The top-left image is the query image, and the similar image returned include the query image itself.)

Figure 5. An example of image retrieval using our algorithm on the Corel-10k dataset. (The top-left image is the query image, and the similar image returned include the query image itself.)

Conclusions

In an image retrieval system, the CNN-based retrieval system often has a better query experience. The returned results of such system are in line with user’s psychological needs. In order to better apply the CNN features to image retrieval and enhance the retrieval performance, we fuse different kinds of CNN features for image retrieval by the weighted average of features similarity and the product rule. Firstly, three different CNN features of images are extracted. Then the weighted feature similarities between the query image and database image are computed by product rule. Finally, the images with top-n highest scores are returned as the retrieval result. Extensive experiments on two publicly available datasets have shown that our algorithm has obvious advantages over that based on individual CNN features and other current retrieval algorithms.

Acknowledgement

This research is supported by the National Natural Science Foundation of China under grants 61572341.

References


